Applied & Translational Genomics 4 (2015) 10–13

Contents lists available at ScienceDirect

Applied & Translational Genomics

journal homepage: www.elsevier.com/locate/atg

Harnessing big data for precision medicine: A panel of experts elucidates the data challenges and proposes key strategic decisions points

Carol Isaacson Barash *, Keith O. Elliston, W. Andrew Faucett, Jonathan Hirsch, Gauri Naik, Alice Rathjen, Grant Wood

A B S T R A C T

A group of disparate translational bioinformatics experts convened at the 6th Annual Precision Medicine Partnership Meeting, October 29–30, 2014 to discuss big data challenges and key strategic decisions needed to advance precision medicine, emerging solutions, and the anticipated path to success. This article reports the panel discussion.

Article

A group of domain experts representing translational bioinformatics companies, precision medical IT for health care institutions and leading health care organizations convened at the 6th Annual Precision Medicine Partnership Meeting, October 29–30, 2014 to discuss big data challenges and key strategic decisions needed to advance precision medicine, emerging solutions, and the anticipated path to success.

Panelists:

- Keith O. Elliston, Ph.D., Chief Executive Officer, TranSMART Foundation
- W. Andrew Faucett, MS, LGC, Director of Policy & Education, Office of the Chief Scientific Officer, Geisinger Health System
- Jonathan Hirsch, Founder & President, Syapse
- Gauri Naik, Ph.D. Chief Scientific Officer, Optra, Inc.
- Alice Rathjen, Founder & Chief Executive Officer, DNA Guide
- Grant Wood, Senior IT Strategist, Clinical Genetics Institute, Intermountain Health Care
- Moderator: Carol Isaacson Barash, Ph.D., Managing Partner, Helix Health Advisors

Defining the big data problem in precision medicine

We tend to think of data as bytes of information and perhaps lose sight of the fact that the source of big data in precision medicine is the human body. Data sources, in other words, have normative beliefs and values about the accessibility and use of their information. The data volume problem starts with the roughly 20,000 genes in each of our bodies, the enormous number of variants within each of those genes, within each of those organs and within each of those cells, and also clinical chemistry, imaging, epigenetic, molecular profiling, tumor profiling, and omics data.

Experts agreed that high volume is one vector of the data problem but the other is complexity. They shared some emerging success strategies that are worth noting. Data storage is transitioning from warehouses to the cloud now that people are beginning to be convinced of cloud security. Geisinger, for example, is storing research exomes in the cloud. Given that two years ago healthcare institutions were deeply dubious about moving away from storage in data warehouses, experts saw this transition as a tremendous progress.

Elliston: Advancing precision medicine requires a different trajectory than has been historically the case. Discovery, development and application require patient centric data to move from the clinic to the research environment and back. The challenge is how to move high volumes of complex data in this loop. The historical practice of people developing, purchasing and implementing large enterprise platforms isn't working in precision medicine. Instead collaboration, crowd sourcing, and open sourcing are the way things are going. Open Clinica and I2B2 are good examples of open source platforms that integrate disparate types of data, but the problem is that they don't integrate molecular profiling and other types of high-dimensional research data. TranSMART integrates clinical information with research data to bring all the necessary information together for the discovery of biomarkers, diagnostics and therapeutics. Patient centered research programs with open collaborations enable people to work together to achieve the goals of precision medicine.

Faucett: Only 2–5% of sequencing data is currently actionable. Discovering which medicines will make a difference for particular patients to represent clinical success points, requires identifying which patients and what a mutation along with a series of other factors. Moving back

* Corresponding author.
E-mail address: cibarash@helixhealthadvisors.com (C.I. Barash).

http://dx.doi.org/10.1016/j.atg.2015.02.002

and forth between clinical and research environments and having access to both sets of data are critical.

Rathjen: How the web evolves, and how nation's decide to regulate patient data will significantly impact what data is available to whom to and thus impact what we can know.

Hirsch: We don't yet have a big data problem in precision medicine. We have a complex data problem. No health care institution has anywhere near the volume of data that companies like Facebook and Google have. The data problems that healthcare currently has are organizational and political. Many are reluctant to adopt new software solutions, and so the industry's challenge is how to change legacy attitudes and embrace new solutions.

Naik: Life science, IT and biopharmaceutical companies are all very interested in mining biodata for in silico drug discovery and drug repurposing purposes. For this reason they are adopting integrated approaches that pair genotypic and phenotypic data with analytics. Optra Systems, for example, is partnering with the IBM Watson consortium to develop a novel NGS platform that will integrate data from different instruments. Inputting data from public and private repositories into the Watson engine, applying a deep quality assurance and analytics is generating results displayed in visualization platforms for evidence-based medicine. The importance of this approach is that it not only integrates genomics into clinical practice but it gives clinicians the evidenced based proof that a particular line of therapy or action is the best.

Wood: One area of data warehousing people are working on is figuring out the types of architectures we need to store not just one genome per person but many "omes" per person. These data sets will need to be linked to the EHR. In the health care world there haven't been resources to attack these types of problems yet but people are starting to plan what is needed.

Problems in data quality and interpretation

Experts stressed that we need a hierarchy to measure data quality to make sure that the most accurate data is the actual data for the read. Data quality is a huge problem because there isn't a gold standard omic technology yet and we're using different technologies now so multiple data points are now the norm.

Experts agreed that interpretation is based on research data that is not well phenotyped. They further agreed that no single lab can interpret an exome or genome. The variance between lab results is emblematic. They agreed that laboratories need to work together to call pathogenicity. Partnerships between laboratories and provider organizations enable the pairing of laboratory reports with clinical outcomes data which is needed to improve interpretation. Accurate interpretation is limited by the inaccessibility of some data sets. Currently ClinGen (www.clinicalgenome.org) and the Global Alliance for Genomic Health are trying to change this.

Elliston: As a systems biologist, I have found that variant interpretation is extremely challenging even when looking at the relatively simple case of dominant variants with full penetrance. When we find the same mutation in three different people linked to the phenotype then we are able to call it and this is a relatively low bar. But when you are looking for the same mutation in non-descended lineages, which is very rare, you may find mutations in the same gene and you may find recessive mutations. The complex system in which genetic variants interact with each other in the same genome, in the same organ, and in the same cells makes interpretation very challenging even for a single gene with a single outcome. For complex diseases, such as neurodegenerative diseases, the challenge is far greater. ***You can find over 300 genes with dominant or recessive inheritance that cause a particular neurodegenerative disease phenotype, with thousands of variants amongst those genes that are incredibly hard to interpret. Even for a single gene disorder with a functional polymorphism, like Huntington Disease, we've no treatments today, twenty years after having identified the gene. So while correctly annotated variants are useful for managing clinical outcomes, they may or may not be useful for identifying therapeutics, which is why research laboratories and clinics need to work collaboratively. Moreover, we need to be mindful of the fact that gene variants are called on the basis of specific standards, but these standards change over time further complicating the problem. In the future we will need to compare, say, the reference genome GRCh35 with version GRCh37 which will then require a standard for comparison. Further, we need to think about what we store and why and how it will be used. We've seen that variants are often called without tracking which reference genome they are called against. Simply storing variance alone does not enable us to compare one variant against the next variant so we need to be careful about the provenance of the data.

Hirsch: Rapid progress is being made on variant interpretation. Laboratories and provider organizations together can advance variant interpretation, through the use of aggregated data including clinical evidence. Healthcare provider systems are uniquely suited to advance variant interpretation because they can pair laboratory reports with clinical space history and outcomes data, which connects lab findings to variant interpretation. Although independent molecular testing labs and interpretation companies are also trying to do this, they do not have access to the clinical evidence needed to confidently interpret the clinical impact of genomic data. Interpreting based on reading published literature is inherently flawed, due to the small sample sizes of most studies. Population-scale variant interpretation will require population-scale clinical studies. We are encourage by the progress that the Global Alliance for Genomics and Health has made in the past year, but much more work needs to be done. We at Syapse are helping lead that charge though our work on population-scale precision medicine with institutions like Intermountain Healthcare.

Naik: Data curation plays a significant role in accurate variant interpretation and its clinical significance. From this perspective many tools, like biological natural language processing, are important to achieving high quality results. These tools can curate 70–80% of the data semi-automatically and accurately. Once more such tools are mainstreamed the authenticity of curation will be improved.

Wood: On the one hand, it's great that new start-ups are selling interpretations. However, caution is needed because these services can't necessarily be mapped onto a healthcare institutions infrastructure. To be useful, we need to answer several questions; namely what tools physicians need to interpret and how we do fit that information into daily clinical workflow. In other words, we need to look at how interpretation should fit into the EHRs. We also need to decide whether our knowledge around interpretation is settled science or still evolving. To give you an example, I'm working on developing an IT infrastructure to house sequencing information and link it to the EHRs. When we have large numbers of patients whose sequencing information in the EHRs, we can then go back and validate where the interpretation came from and determine whether it's correct or not. This capability will develop over time. But given the volume of data being generated we need to solve this problem quickly. We also need to solve the standardization problem soon to avoid physicians ordering repeat tests because they don't trust a lab result.

Faucett: All clinical laboratories need their data to be standardized and need to input phenotypic data to go along with it. The problem now is that interpretation is often based on research data and research data isn't phenotyped well. Further, no single lab can develop the internal database to interpret an exome or a genome. The variance within and across lab results is telling. Some labs take a lot of time to call a variant and have a low variant-of-uncertain-significance rate, while others don't and still others are unwilling to make a call. ClinVar and ClinGen are developing a system to capture the evidence behind interpretation, which will permit you to see if an interpretation was reported by a single lab, or two labs and if the latter case, whether it was reported consistently. It can also enable you to see whether an interpretation was reported by more than two labs, in which case you

can see whether someone did an independent review. We need more evidence-based guidelines for interpretation so we can determine whether this is the variant we're calling pathogenic and what our level of certainty is that our call is correct. In sum, labs need to work more closely together.

Rathjen: A separate challenge is the fact that the information changes so fast and this speed will continue for a long time.

Machine learning, curation and knowledge generation

Experts agreed that big data in precision medicine will become not only more coordinated but that interpretation and validation require machine learning. But the field is in its infancy.

Elliston: We've learned that trying to reproduce research results requires considerable work, and the results are not always reproducible. Many believe that these irreproducible results are just bad experiments. The problem is that rather than cherry picking results, we need to learn directly from all of the data in a hypothesis free mode. This is why machine learning will be an invaluable tool for biology. In biology, context matters, and the confluence of factors in dynamic interplay can be accurately interpreted by intelligent systems that look for the relationships between variables. The challenge is how to do this in the right ways with the right data and then to understand what we've learned. This is a new area of research, and one that is going to be growing very rapidly in the healthcare space.

Rathjen: I have a vision of self-organizing genomes with real time consent that enables big data to be coordinated. Much of the literature is based on small samples, and not replicable. That alone is not motivation to sort through studies to decide which are good or bad, given that we can create something so much more robust in the future.

Hirsch: Machine learning may be the future, but machine learning on *what*? Presently, health IT is consumed by meeting Meaningful Use 2, and many more requirements are coming. Precision medicine, complex clinical decision support, population health, machine learning, and learning health systems are beyond the scope of what EHRs can handle. In order to implement something as simple as enterprise-wide pharmacogenomics clinical decision support interfacing with the EMR, Syapse software must do all of the complex genomics rules calculations and serve the simplified decision support into the EHR, because the EHR is not capable of handling the necessary rules logic. It will be interesting to see how EHR vendors decide to handle genomic data storage and integrate the complex logic needed to drive decisions in the clinical genomic space.

Naik: The machine learning that exists today is a baby step towards the future and its not just the content produced that is important, it's the context.

Faucett: I would say we need to continue to work on using machine learning for curation but machine learning has much to learn before it can effectively be used to combine clinical curation and expert mining of phenotypic data. We already have examples of people known to have pathogenic sequences but not disease. If you look at their clinical history they have problems but have not been diagnosed so they're disease free and don't fit the classic phenotype. This shows that we need to go back and forth between clinical and research data if we're going to correctly annotate variants. I view machine learning as a screening tool that helps us identify what we need to look at further. It may also help us identify a clinical significance so that we don't need to curate *every* variant, at least for the near future.

Wood: I look at machine learning as an intelligent capability built into the EHRs. To date the only intelligence we've built into the EHR is decision support that's been mainly risk algorithms around clinical data. The need to clinically use omic data creates an opportunity to take this to the next level. At the Institute of Medicine Collaborative we are working with EHR vendors to determine what they need to start storing genomic data in the EHR in order to create a standard which all vendors can use. The HL7 group established standards several years ago so the IOM collaborative has established 22 minimum data elements that need to reside in the HER. We are in the process of mapping these requirements onto the HL7 standards.

We can't store genomic data the way we store lab results in the EHRs. If all we needed was a single gene we could use the same standard. The structure I envision enables a physician who needs genomic data in order to make the best decision for a patient, whether it be to diagnose a disease, identify disease risks or prescribe the right drug. A physician would be able to link to a repository containing the patient's omic data, find the needed information and link back out to the EHR, or enter the genomic information in the EHR and enter the decision. For example, if a physician is assessing a patient's risk for colon cancer they may want to see if the patient has KRAS mutations, so they link out to the repository, find the answer, and put it into the EHRs. Some say this is too much automated intelligence because it's a mistake to exclude clinical geneticists in this process.

Data access & consent

Elliston: Although the patient consent landscape is changing, these consents are currently a limitation to realizing the value of patient data. In personalized medicine we have to track which consents patients have given and for which analyses in order to know what we can and cannot do with their data. If you look outside this domain to see how consent is handled with other types of big data, the situation is entirely different. In Facebook, your consent was probably tacit; a click-thru terms of service agreement that you likely did not even read or really agree to. In this case, sharing your data doesn't benefit you or the community in the least. It only benefits the vendor and its customers (advertisers). With healthcare data, you are not presented with a click-thru consent option and most people don't realize that their data is being restricted. I think this is all upside down. We should be able to restrict how our data is used on Facebook and share our healthcare data so that people can work on treating and curing disease. This is of true social benefit. I think this consent issue is one of the major socio-economic barriers to moving genomic medicine to the next level.

Hirsch: When we implement our software to support a precision medicine program in a healthcare organization, the healthcare organization typically puts in place a consent infrastructure that allows data to be used for research and clinical purposes. Though there is tremendous variability across our customer set on this front, and one of the challenges our industry faces is variability of informed consents and IRBs. Regarding patient motivations, our customers have found that patients are willing to share data if it helps advance research into diseases that impact them or their families.

Faucett: I think most people are willing to consent to broad use of their data, but they want to know about it. Some aren't and consent needs to have provisions that permit these people to opt-out. The important point is that historically consent is implied and the process is not transparent. Consenting does cost a bit but we need to make it explicit in our practice. Research supports the notion that most people will share their data. People want to be sure that the right guardianship is in place and for this reason it's important that we are transparent because transparency builds trust and trust promotes data sharing. The public knows we can't guarantee the privacy of their data but that we can do our best to protect it and we need to be clear about how we protect it. This is particularly true since every day the public is seeing privacy breaches.

Moving towards patient centric care

Experts agreed that patient centric care requires patients owning and controlling their data and that health care institutions will be moving in this direction. Patient centric care also will involve moving between research and clinical data. Patients, for example, are likely to ask their physicians for their clinical opinions about sequences

generated in research and so data needs to be free flowing. It further means that patients will contribute to their data and take an active role in better understanding their health challenges and improving their outcomes. In the future people will manage their care on their mobile devices so the data needs to be accessible and linked to disparate entities. As patients contribute to their data, we will begin to focus on families and not just individuals.

In sum, the panel agreed that some important shifts have already occurred. Data storage is transitioning from warehouses to the cloud as organizations are increasingly convinced of cloud security. Organizations have moved from being opposed to open source solutions to embracing them and collaborative platforms. Data integration is increasing because we are adopting integrated approaches. Precision medicine will adapt the big data volume solutions being pioneered by the companies, like Google, Amazon, Microsoft or Rackspace, that are wrestling with how to design software on cloud. The data complexity problem, however, will be Precision Medicine's challenge to solve. In the future we will always use mined data and we can move forward to create a learning health care system if we all work together.